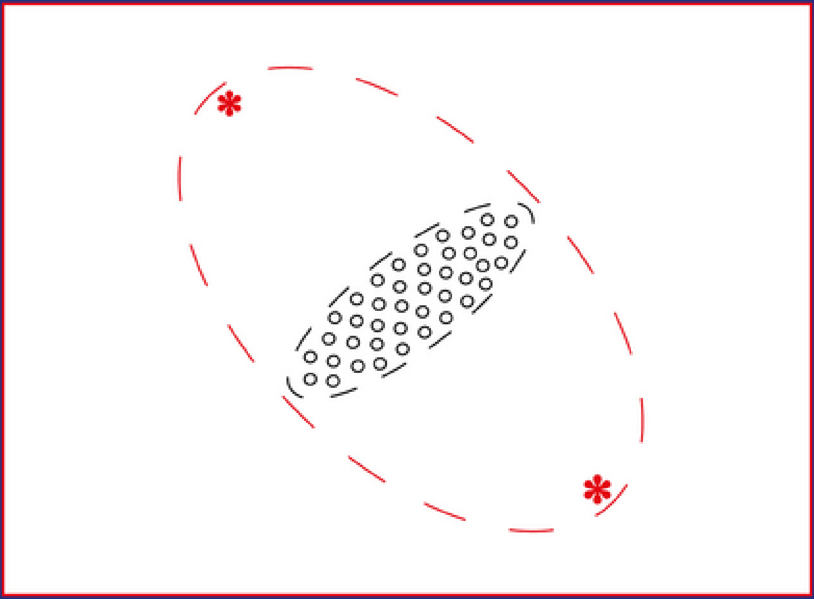
Wiley Series in Probability and Statistics

# Robust Correlation
## Theory and Applications

Georgy L. Shevlyakov • Hannu Oja



with website

WILEY

# Robust Correlation

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

A complete list of the titles in this series appears at the end of this volume.

# Robust Correlation

## Theory and Applications

**Georgy L. Shevlyakov**

*Peter the Great Saint-Petersburg Polytechnic University, Russia*

**Hannu Oja**

*University of Turku, Finland*

WILEY

**Registered office**
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. It is sold on the understanding that the publisher is not engaged in rendering professional services and neither the publisher nor the author shall be liable for damages arising herefrom. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

***Library of Congress Cataloging-in-Publication Data***

Names: Shevlyakov, Georgy L. | Oja, Hannu.
Title: Robust correlation : theory and applications / Georgy L. Shevlyakov,
Peter The Great Saint-Petersburg Polytechnic University, Russia, Hannu
Oja, University Of Turku, Finland.
Description: Chichester, West Sussex : John Wiley & Sons, Inc., 2016. |
Series: Wiley series in probability and statistics | Includes
bibliographical references and index.
Identifiers: LCCN 2016017308 (print) | LCCN 2016020693 (ebook) | ISBN
9781118493458 (cloth) | ISBN 9781119264538 (pdf) | ISBN 9781119264491
(epub)
Subjects: LCSH: Correlation (Statistics) | Mathematical statistics.
Classification: LCC QA278.2 .S4975 2016 (print) | LCC QA278.2 (ebook) | DDC
519.5/37–dc23
LC record available at https://lccn.loc.gov/2016017308

A catalogue record for this book is available from the British Library.

Set in 10/12pt, Times by SPi Global, Chennai, India.

1 2016

*To our families*

# Contents

# Preface

Robust statistics as a branch of mathematical statistics appeared due to the seminal works of John W. Tukey (1960), Peter J. Huber (1964), and Frank R. Hampel (1968). It has been intensively developed since the sixties of the last century and is definitely formed by the present. The term “robust” (Latin: strong, sturdy, tough, vigorous) as applied to statistical procedures was proposed by George E.P. Box (1953).

The principal reason for research in this field of statistics is of a general mathematical nature. Optimality (accuracy) and stability (reliability) are the mutually complementary characteristics of many mathematical procedures. It is well-known that the performance of optimal procedures is, as a rule, rather sensitive to “small” perturbations of prior assumptions. In mathematical statistics, the classical example of such an unstable optimal procedure is given by the least squares method: its performance may become disastrously poor under small deviations from normality.

Roughly speaking, robustness means stability of statistical inference under the departures from the accepted distribution models. Since the term “stability” is generally overloaded in mathematics, the term “robustness” may be regarded as its synonym.

Peter J. Huber and Frank R. Hampel contributed much to robust statistics: they proposed and developed two principal approaches to robustness, namely, the minimax approach and the approach based on influence functions, which were applied to almost all areas of statistics: robust estimation of location, scale, regression, and multivariate model parameters, as well as to robust hypothesis testing. It is remarkable that although robust statistics involves mathematically highly refined asymptotic tools, nevertheless robust methods show a satisfactory performance in small samples, being quite useful in applications.

The main topic of our book is robust correlation. Correlation analysis is widely used in multivariate statistics and data analysis: computing correlation and covariance matrices is both an initial and a basic step in most procedures of multivariate statistics, for example, in principal component analysis, factor and discriminant analysis, detection of multivariate outliers, etc.

Our work represents new results generally related to robust correlation and data analysis technologies, with definite accents both on theoretical aspects and practical needs of data processing: we have written the book to be accessible both to the users of statistical methods as well as to professional statisticians. However, the mathematical background requires the basics of calculus, linear algebra, and mathematical statistics.

Chapter 1 is an introduction into the book, providing historical aspects of the origin and development of the notion “correlation” in science as well as ontological remarks on the subject of statistics and data processing. Chapter 2 delivers a survey of the classical measures of correlation aimed most at estimating linear dependencies.

Chapter 3 represents Huber’s and Hampel’s principal approaches to robustness in mathematical statistics, with novel additions to them, namely, a stable estimation approach and an essay on robustness versus Gaussianity, the latter of which could be helpful for students and their teachers. Except for a few paragraphs on the application of Huber’s minimax approach to distribution classes of a non-neighborhood nature, Chapters 1 to 3 are accessible to a wide reader audience.

Chapters 4 to 8 comprise the core of the book, which contains most of the new theoretical and experimental (Monte Carlo) results. Chapter 4 treats the problems of robust estimation of a scale parameter, and the obtained results are used in Chapter 5 for the design of highly robust and efficient estimates of a correlation coefficient including robust minimax (in the Huber sense) estimates. Chapter 6 provides an overview of classical multivariate correlation measures and inference tools based on the covariance and correlation matrix. Chapter 7 deals with robust correlation measures and inference tools that are based on various robust covariance matrix functionals and estimates; in particular, robust versions of principal component and canonical correlation analysis are given. Chapter 8 comprises correlation measures and inference tools based on various concepts of univariate and multivariate signs and ranks.

Chapters 9 to 11 are devoted to the applications of the aforementioned robust estimates of correlation, as well as of location and scale, to different problems of statistical data and signal analysis, with a few examples of real-life data and signal processing. Chapter 9 is confined to the applications to exploratory data analysis and its technologies, mostly treating an important problem of detection of outliers in the data. Chapter 10 outlines a few novel approaches to robust estimation of time series power spectra: although the obtained results are preliminary, they are profitable, deserving a further thorough study. In Chapter 11, various problems of robust signal detection are posed and treated, in the solution of which the Huber’s minimax and stable approaches to robust detection are successfully exploited.

Chapter 12 outlines several open problems in robust multivariate analysis and its applications.

From the aforementioned it follows that there are two main blocks of the book: Chapters 1 to 3 and 9 to 11 aim at the applied statistician and statistics user audience, while Chapters 4 to 8 focus on the theoretical aspects of robust correlation.

Most of the contents of the book, namely Chapters 1 to 5 and 9 to 11, have been written by the first author. The second author contributed Chapters 6 to 8 on general multivariate analysis.

# Acknowledgements

John W. Tukey, Peter J. Huber, Frank R. Hampel, Elvezio M. Ronchetti, and Peter J. Rousseeuw have essentially influenced, directly or indirectly, our views on robustness and data analysis.

The first author is deeply grateful to his teachers and colleagues for their helpful and constructive discussions, namely, to Igor B. Chelpanov, Peter Filzmoser, Eugene P. Gilbo, Jana Jureckova, Abram M. Kagan, Vladimir Ya. Katkovnik, Yuriy S. Kharin, Kiseon Kim, Lev B. Klebanov, Stephan Morgenthaler, Yakov Yu. Nikitin, Boris T. Polyak, Alexander M. Shurygin, and Nikita O. Vilchevski.

Some results presented in Chapters 4, 5, and 9 to 11 by the first author are based on the Ph.D. and M.Sc. dissertations of his former students, including Kliton Andrea, JinTae Park, Pavel Smirnov, Galina Lavrentyeva, Nickolay Lyubomishchenko, and Nikita Vassilevskiy—we would like to thank them.

Research on multivariate analysis reported by the second author is to some degree based on the thesis works of his several ex-students, including Jyrki Möttönen, Samuli Visuri, Esa Ollila, Sara Taskinen, Seija Sirkiä, and Klaus Nordhausen. We wish to thank them all. The second author is naturally also indebted to many colleagues and coauthors for valuable discussions and express his sincere thanks for discussions and cooperation in this specific research area with Christopher Croux, Tom Hettmansperger, Annaliisa Kankainen, Visa Koivunen, Ron Randles, Bob Serfling, and Dave Tyler.

We are also grateful to Igor Bezdvornyh and Maksim Sovetnikov for their technical help in the preparation of the manuscript.

Finally, we wish to thank our wives, Elena and Ritva, for their patience, support, and understanding.

# About the companion website

Don’t forget to visit the companion website for this book:

**www.wiley.com/go/Shevlyakov/Robust**

There you will find valuable material designed to enhance your learning, including:

- Datasets
- R codes

Scan this QR code to visit the companion website.

# 1

# Introduction

This book is most about correlation, association and partially about regression, i.e., about those areas of science where the dependencies between random variables that mathematically describe the relations between observed phenomena and associated with them features are studied. Evidently, these concepts and terms firstly appeared in applied sciences, not in mathematics. Below we briefly overview the historical aspects of the considered concepts.

## 1.1 Historical Remarks

The word "*correlation*" is of late Latin origin meaning "*association*", "*connection*", "*correspondence*", "*interdependence*", "*relationship*", but relationship not in the conventional for that time deterministic functional form.

The term "*correlation*" was introduced into science by a French naturalist Georges Cuvier (1769–1832), one of the major figures in natural sciences in the early 19th century, who had founded paleontology and comparative anatomy. Cuvier discovered and studied the relationships between the parts of animals, between the structure of animals and their mode of existence, between the species of animals and plants, and many others. This experience made him establish the general principles of "*the correlation of parts*" and of "*the functional correlation*" (Rudwick 1997):

*Today comparative anatomy has reached such a point of perfection that, after inspecting a single bone, one can often determine the class, and sometimes even the genus of the animal to which it belonged, above all if that bone belonged to the head or the limbs. … This is because the number, direction, and shape of the bones that compose each part of an animal's body are always in a necessary relation to all the*

*Robust Correlation: Theory and Applications*, First Edition. Georgy L. Shevlyakov and Hannu Oja.

Companion Website: www.wiley.com/go/Shevlyakov/Robust

*other parts, in such a way that – up to a point – one can infer the whole from any one of them and vice versa.*

From Cuvier to Galton, correlation had been understood as a qualitatively described relationship, not deterministic but of a statistical nature, however observed at that time within a rather narrow area of phenomena.

The notion of *regression* is connected with the great names of Laplace, Legendre, Gauss, and Galton (1885), who coined this term. Laplace (1799) was the first to propose a method for processing the astronomical data, namely, the least absolute values method. Legendre (1805) and Gauss (1809) independently of each other introduced the least squares method.

Francis Galton (1822–1911), a British anthropologist, biologist, psychologist, and meteorologist, understood that correlation is the interrelationship in average between any random variables (Galton 1888):

*Two variable organs are said to be co-related when the variation of the one is accompanied on the average by more or less variation of the other, and in the same direction. … It is easy to see that co-relation must be the consequence of the variations of the two organs being partly due to common cause. … If they were in no respect due to common causes, the co-relation would be nil.*

Correlation analysis (this term also was coined by Galton) deals with estimation of the value of correlation by number indexes or coefficients.

Similarly to Cuvier, Galton introduced regression dependence observing live nature, in particular, processing the heredity and sweet peas data (Galton 1894). Regression characterizes the correlation dependence between random variables functionally in average. Studying the sizes of sweet peas beans, he noticed that the offspring seeds did not reveal the tendency to reproduce the size of their parents being closer to the population mean than them. Namely, the seeds were smaller than their parents in the case of large parent sizes, and vice versa. Galton called this dependence regression, for the reverse changes had been observed: firstly, he used the term "*the law of reversion*". Further studies showed that on average the offspring regression to the population mean was proportional to the parent deviations from it – this allowed the observed dependence to be described using the linear function. The similar linear regression is described by Galton as a result of processing the heights of 930 adult children and their 205 parents (Galton 1894).

The term "*regression*" became popular, and now it is used in the case of functional dependencies in average between any random variables. Using modern terminology, we may say that Galton considered the slope $r$ of the simple linear regression line as a measure of correlation (Galton 1888):

*Let $y =$ the deviation of the subject [in units of the probably error, $Q$], whichever of the two variables may be taken in that capacity; and let $x_1, x_2, x_3,$ & c., be the corresponding deviations of the relative, and let the mean of these be $X$. Then we find: (1) that $y = r\,X$ for all values of $y$; (2) that $r$ is the same, whichever of the two variables is taken for the subject; (3) that $r$ is always less than 1; (4) that $r$ measures the closeness of co-relation.*

Now we briefly comment on the above-mentioned properties (1)–(4): the first is just the simple linear regression equation between the standardized variables $X$ and $y$; the second means that *the co-relation* $r$ is symmetric with regard to the variables $X$ and $y$; the third and fourth show that Galton had not yet recognized the idea of negative correlation: stating that $r$ could not be greater than 1, he evidently understood $r$ as a positive measure of *"co-relation"*. Originally $r$ stood for the regression slope, and that is really so for the standardized variables; Galton perceived the correlation coefficient as a scale invariant regression slope.

Galton contributed much to science studying the problems of heredity of qualitative and quantitative features. They were numerically examined by Galton on the basis of the concept of correlation. Until the present, the data on demography, heredity, and sociology collected by Galton with the corresponding numerical examples of correlations computed are used.

Karl Pearson (1857–1936), a British mathematician, statistician, biologist, and philosopher, had written out the explicit formulas for the population product-moment correlation coefficient (Pearson 1895)

$$\rho = \rho(X, Y) = \frac{cov(X, Y)}{[var(X)\ var(Y)]^{1/2}} \tag{1.1}$$

and its sample version

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{1/2}} \tag{1.2}$$

(here $\bar{x}$ and $\bar{y}$ are the sample means of the observations $\{x_i\}$ and $\{y_i\}$ of random variables $X$ and $Y$). However, Pearson did not definitely distinguish the population and sample versions of the correlation coefficient, as it is commonly done at present.

Thus, on the one hand, the sample correlation coefficient $r$ is a statistical counterpart of the correlation coefficient $\rho$ of a bivariate distribution, where $var(X)$, $var(Y)$, and $cov(X, Y)$ are the variances and the covariance of the random variables $X$ and $Y$, respectively.

On the other hand, it is an efficient maximum likelihood estimate of the correlation coefficient $\rho$ of the bivariate normal distribution (Kendall and Stuart 1963) with density

$$N(x, y; \mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\right.$$

$$\left.\times \left[\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\}, \tag{1.3}$$

where $\mu_X = E(X)$, $\mu_Y = E(Y)$, $\sigma_X^2 = var(X)$, $\sigma_Y^2 = var(Y)$.

Galton (1888) derived the bivariate normal distribution (1.3), and he was the first who used it to scatter the frequencies of children's stature and parents' stature. Pearson noted that "in 1888 Galton had completed the theory of bivariate normal correlation" (Pearson 1920).

Like Galton, Auguste Bravais (1846), a French naval officer and astronomer, came very near to the definition (1.1) when he called one parameter of the bivariate normal distribution "une correlation", but he did not recognize it as a measure of the interrelationship between variables. However, "his work in Pearson's hands proved useful in framing formal approaches in those areas" (Stigler 1986).

Pearson's formulas (1.1) and (1.2) proved to be fruitful for studying dependencies: correlation analysis and most of multivariate statistical analysis tools are based on the pair-wise Pearson correlations; we may also add the correlation and spectral theories of stochastic processes, etc.

Since the time Pearson introduced the sample correlation coefficient (1.2), many other measures of correlation have been used aiming at estimation of the closeness of interrelationship (the coefficients of association, determination, contingency, etc.). Some of them were proposed by Karl Pearson (1920).

It would not be out of place to note the contributions to correlation analysis of the other British statisticians.

Ronald Fisher (1890–1962) is one of the creators of mathematical statistics. In particular, he is the originator of the analysis of variance and together with Karl Pearson he stands at the beginning of the theory of hypothesis testing. He introduced the notion of a sufficient statistic and proposed the maximum likelihood method (Fisher 1922). Fisher also payed much attention to correlation analysis: his tools for verifying the significance of correlation under the normal law are used until now.

George Yule (1871–1951) is a prominent statistician of the first half of the 20th century. He contributed much to the statistical theories of regression, correlation (Yule's coefficient of contingency between random events), and spectral analysis.

Maurice Kendall (1907–1983) is one of the creators of nonparametric statistics, in particular, of the nonparametric correlation analysis (the Kendall $\tau$-rank correlation) (Kendall 1938). It is noteworthy that he is the coauthor of the classical course in mathematical statistics (Kendall and Stuart 1962, 1963, 1968).

In what follows, we represent their contributions to correlation analysis in more detail.

## 1.2 Ontological Remarks

Our personal research experience in applied statistics and real-life data analysis is relatively broad and long. It is concerned with the problems of data processing in medicine (cardiology and ophthalmology), biology (genetics), economics and finances (financial mathematics), industry (mechanical engineering, energetics, and material science), and analysis of semantic data and informatics (information retrieval from big data). Besides and due to those problems, we have been working in theoretical statistics, most in robust and nonparametric statistics, as well as in multivariate statistics and time series analysis. Now we briefly outline our vision of the topic of this book to indicate its place in the general context of statistical data analysis with its philosophy and ideological environment.

*The reader should only remember that any classification is a convention, such are the forthcoming ones.*

### 1.2.1 Forms of data representation

The customary forms of data representation are as follows (Shevlyakov and Vilchevski 2002, 2011):

- as a sample $\{x_1, \cdots, x_n\}$ of real numbers $x_i$ being the most convenient form to deal with;
- as a sample $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ of real-valued vectors $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})^T$ of dimension $p$;
- as an observed realization $x(t)$, $t \in [0, T]$ of a real-valued continuous process (function);
- as a sample of “non-numerical nature” data representing qualitative variables;
- as the semantic type of data (statements, texts, pictures, etc.).

The first three possibilities mostly occur in the natural and technical sciences with the measurement techniques being well developed, clearly defined, and largely standardized. In the social sciences, the last forms are relatively common.

*To summarize: in this book we deal mostly with the first three forms and, partially, with the fourth.*

### 1.2.2 Types of data statistics

The experience of treating various statistical problems shows that practically all of them are solved with the use of only a few qualitatively different types of data statistics. Here we do not discuss how to use them in solving statistical problems: only note that their solutions result in computing some of those statistics, and final decision making essentially depends on their values (Mosteller and Tukey 1977; Tukey 1962).

These data statistics may be classified as follows:

- measures of location (central tendency, mean values),
- measures of scale (spread, dispersion, scatter),

- measures of correlation (interdependence, association),
- measures of extreme values,
- measures of a data distribution shape,
- measures of data spectrum.

*To summarize: in this book we mainly focus on the measures of correlation, however dealing if needed with the other types of data statistics.*

### 1.2.3 Principal aims of statistical data analysis

These aims can be formulated as follows:

*(A1)* compact representation of data,

*(A2)* estimation of model parameters explaining and/or revealing data structure,

*(A3)* prediction.

A human mind cannot efficiently work with large volumes of information, since there exist natural psychological bounds on the perception ability (Miller 1956). Thus it is necessary to provide a compact data output of information for expert analysis: only in this case we may expect a satisfactory final decision. Note that data processing often begins and ends with the first item *(A1)*.

The next step *(A2)* is to propose an explanatory underlying model for the observed data and phenomena. It may be a regression model, or a distribution model, or any other, desirably a low-complexity one: an essentially multiparametric model is usually a "bad" model; nevertheless, we should recall a cute note of George Box: "*All models are wrong, but some of them are useful*" (Box and Draper 1987). However, parametric models are the first to consider and examine.

Finally, the first two aims are only the steps to the last aim *(A3)*: here we have to state that this aim remains a main challenge to statistics and to science as a whole.

*To summarize: in this book we pursue aims (A1) and (A2).*

### 1.2.4 Prior information about data distributions and related approaches to statistical data analysis

The need for stability in statistical inference directly leads to the use of robust statistical methods. It may be roughly stated that, with respect to the level of prior information about underlying data distributions, robust statistical methods occupy the intermediate place between classical parametric and nonparametric methods.

In parametric statistics, the shape of an underlying data distribution is assumed known up to the values of unknown parameters. In nonparametric statistics, it is supposed that the underlying data distribution belongs to some sufficiently “wide” class of distributions (continuous, symmetric, etc.). In robust statistics, at least within Huber’s minimax approach (Huber 1964), we also consider distribution classes but with more detailed information about the underlying distribution, say, in the form of a neighborhood of the normal distribution. The latter peculiarity allows the efficiency of robust procedures to be raised as compared with nonparametric methods, simultaneously retaining their high stability.

At present, there exist two main approaches in robustness:

- Huber’s minimax approach — quantitative robustness (Huber 1981; Huber and Ronchetti 2009).

- Hampel’s approach based on influence functions — qualitative robustness (Hampel 1968; Hampel *et al.* 1986).

In Chapter 3, we describe these approaches in detail. Now we classify the existing approaches in statistics with respect to the level of prior information about the underlying data distribution $F(x;\theta)$ in the case of point parameter estimation:

- A given data distribution $F(x;\theta)$ with a random parameter $\theta$ — the Bayesian statistics (Berger 1985; Bernardo and Smith 1994; Jaynes 2003).

- A given data distribution $F(x;\theta)$ with an unknown parameter $\theta$ — the classical parametric statistics (Fisher 1922; Kendall and Stuart 1963).

- A data distribution $F(x;\theta)$ with an unknown parameter $\theta$ belongs to a distribution class $\mathcal{F}$, usually a neighborhood of a given distribution, e.g., normal — the robust statistics (Hampel *et al.* 1986; Huber 1981; Kolmogorov 1931; Tukey 1960).

- A data distribution $F(x;\theta)$ with an unknown parameter $\theta$ belongs to some general distribution class $\mathcal{F}$ — the classical nonparametric statistics (Hettmansperger and McKean 1998; Kendall and Stuart 1963; Wasserman 2007).

- A data distribution $F(x;\theta)$ does not exist in the case of unique samples and frequency instability — the probability-free approaches to data analysis: fuzzy (Zadeh 1975), exploratory (Bock and Diday 2000; Tukey 1977), interval probability (Kuznetsov 1991; Walley 1990), logical-algebraic, geometrical (Billard and Diday 2003; Diday 1972).

Note that the upper and lower levels of this hierarchy, namely the Bayesian and the probability-free approaches, are being intensively developed at present.

*To summarize: in this book we mainly use Huber's and Hampel's robust approaches to statistical data analysis.*

## References

Berger JO 1985 *Statistical Decision Theory and Bayesian Analysis,* Springer.

Bernardo JM and Smith AFM 1994 *Bayesian Theory,* Wiley.

Billard L and Diday E 2003 From the statistics of data to the statistics of knowledge: symbolic data analysis. *J. Amer. Statist. Assoc.* **98**, 991–999.

Bock HH and Diday E (eds) 2000 *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer.

Box GEP and Draper NR 1987 *Empirical Model-Building and Response Surfaces*, Wiley.

Bravais A 1846 Analyse mathématique sur les probabilités des erreurs de situation d'un point. *Mémoires presents par divers savants l'Académie des Sciences de l'Institut de France. Sciences Mathématiques et Physiques* **9**, 255–332.

Diday E 1972 Nouvelles Méthodes et Nouveaux Concepts en Classification Automatique et Reconnaissance des Formes. These de doctorat d'état, Univ. Paris IX.

Fisher RA 1922 On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society*, A **222**, 309–368.

Galton F 1885 Regression towards mediocrity in hereditary stature. *Journal of Anthropological Institute* **15**, 246–263.

Galton F 1888 Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London* **45**, 135–145.

Galton F 1894 *Natural Inheritance*, Macmillan, London.

Gauss CF 1809 *Theoria Motus Corporum Celestium*, *Perthes, Hamburg; English translation: Theory of the Motion of the Heavenly Bodies Moving about the Sun in Conic Sections*. New York: Dover, 1963.

Hampel FR 1968 *Contributions to the Theory of Robust Estimation*. PhD thesis, University of California, Berkeley.

Hampel FR, Ronchetti E, Rousseeuw PJ, and Stahel WA 1986 *Robust Statistics. The Approach Based on Influence Functions*, Wiley.

Hettmansperger TP and McKean JW 1998 *Robust Nonparametric Statistical Methods. Kendall's Library of Statistics*, Edward Arnold, London.

Huber PJ 1964 Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73–101.

Huber PJ 1981 *Robust Statistics*, Wiley.

Huber PJ and Ronchetti E (eds) 2009 *Robust Statistics*, 2nd edn, Wiley.

Jaynes AT 2003 *Probability Theory. The Logic of Science*, Cambridge University Press.

Kendall MG 1938 A new measure of rank correlation. *Biometrika* **30**, 81–89.

Kendall MG and Stuart A 1962 *The Advanced Theory of Statistics. Distribution Theory*, vol. 1, Griffin, London.

Kendall MG and Stuart A 1963 *The Advanced Theory of Statistics. Inference and Relationship*, vol. 2, Griffin, London.

Kendall MG and Stuart A 1968 *The Advanced Theory of Statistics. Design and Analysis, and Time Series*, vol. 3, Griffin, London.

Kolmogorov AN 1931 On the method of median in the theory of errors. *Math. Sbornik* **38**, 47–50.

Kuznetsov VP 1991 *Interval Statistical Models*, Radio i Svyaz, Moscow (in Russian).

Legendre AM 1805 *Nouvelles methods pour la determination des orbits des cometes*, Didot, Paris.

Miller GA 1956 The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* **63**, 81–97.

Mosteller F and Tukey JW 1977 *Data Analysis and Regression*, Addison–Wesley.

Pearson K 1895 Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London* A **186**, 343–414.

Pearson K 1920 Notes on the history of correlations. *Biometrika* **13**, 25–45.

Rudwick MJS 1997 *Georges Cuvier, Fossil Bones, and Geological Catastrophes*, University of Chicago Press.

Shevlyakov GL and Vilchevski NO 2002 *Robustness in Data Analysis: criteria and methods*, VSP, Utrecht.

Shevlyakov GL and Vilchevski NO 2011 *Robustness in Data Analysis*, De Gruyter, Boston.

Stigler SM 1986 *The History of Statistics: The Measurement of Uncertainty before 1900.* Belknap Press/Harvard University Press.

Tukey JW 1960 A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics*. (ed. Olkin I). pp. 448–485. Stanford Univ. Press.

Tukey JW 1962 The future of data analysis. *Ann. Math. Statist.* **33**, 1–67.

Tukey JW 1977 *Exploratory Data Analysis*, Addison–Wesley.

Walley P 1990 *Statistical Reasoning with Imprecise Probabilities*, Chapman & Hall.

Wasserman L 2007 *All of Nonparametric Statistics*, Springer.

Zadeh LA 1975 Fuzzy logic and approximate reasoning. *Synthese* **30**, 407–428.

# 2

# Classical Measures of Correlation

In this chapter we define several conventional measures of correlation, focusing most on Pearson's correlation coefficient and closely related to it constructions, enlist their principal properties and computational peculiarities.

## 2.1 Preliminaries

Here we comment on the requirements that should be imposed on the measures of correlation to distinguish them from the measures of location and scale (Renyi 1959; Schweizer and Wolff 1981).

Let $corr\,(X, Y)$ be a measure of correlation between any random variables $X$ and $Y$. Here we consider both positive–negative correlation

$$-1 \leq corr\,(X, Y) \leq 1$$

and positive correlation

$$0 \leq corr\,(X, Y) \leq 1.$$

It is natural to impose the following requirements on $corr(X, Y)$:

(R1) Symmetry: $corr\,(X, Y) = corr\,(Y, X)$.

(R2) Invariancy to linear transformations of random variables:

$$corr\,(aX + b, cY + d) = \operatorname{sgn}(ac)\; corr\,(X, Y), \quad ac \neq 0.$$

*Robust Correlation: Theory and Applications*, First Edition. Georgy L. Shevlyakov and Hannu Oja.

Companion Website: www.wiley.com/go/Shevlyakov/Robust

(R3) Attainability of the limit values 0, +1, and −1:

(a) for independent $X$ and $Y$, $corr\,(X, Y) = 0$;

(b) $corr\,(X, X) = 1$;

(c) $corr\,(X, aX + b) = \text{sgn}(a)$ for positive-negative correlation.

(R4) Invariancy to strictly monotonic transformations of random variables:

$$corr\,(g(X), h(Y)) = \pm corr\,(X, Y)$$

for strictly monotonic functions $g(X)$ and $h(Y)$.

(R5) $corr\,(g(X), h(X)) = \pm 1$.

Requirement (R1) holds almost for all known measures of correlation, being a natural assumption for correlation analysis as compared to regression analysis when it is not known which variables are dependent and which not.

Requirement (R2) makes a measure of correlation independent of the chosen measures of location and scale, since each of them reflects qualitatively different data characteristics.

Requirements (R3a), (R3b), and (R3c), on the one hand, are merely technical: it is practically and theoretically convenient to deal with a bounded scaleless measure of correlation; on the other hand, they refer to the correspondence of the limit values of $corr\,(X, Y)$ to the limit cases of association between random variables $X$ and $Y$: the relation $corr\,(X, Y) = 0$ may mean independence of $X$ and $Y$ whereas the relation $|corr\,(X, Y)| = 1$ indicates the functional dependence between $X$ and $Y$.

The first three requirements hold for almost all known measures of correlation. This is not so with the relation $corr\,(X, Y) = 0$, which does not guarantee independence of $X$ and $Y$ for several measures of correlation, for example, for Pearson's product-moment, the Spearman rank correlation, and for a few others.

However, this property holds for the maximal correlation coefficient $S(X, Y)$ defined as

$$S(X, Y) = \sup_{g,h} \rho(g(X), h(Y)),$$

where $\rho(X, Y)$ is Pearson's product-moment correlation coefficient (1.1), $g(x)$ and $h(x)$ are Borel-measurable functions such that $\rho(g(X), h(Y))$ has sense (Gebelein 1941). The independence of random variables $X$ and $Y$ also follows from the null value of Sarmanov's correlation coefficient (also called the maximal correlation coefficient) (Sarmanov 1958): in the case of a continuous symmetric bivariate distribution of $X$ and $Y$, its value is reciprocal to the minimal eigenvalue of some integral operator. Apparently, Gebelein's and Sarmanov's correlation coefficients are rather complicated in their usage.

Recently, in Székely *et al.* (2007), a distance correlation has been proposed: its equality to null implies independence, but like Gebelein's and Sarmanov's correlation coefficients, it is much more complicated in computing than classical measures of correlation.

Requirements (R4) and (R5) refer to the rank measures of correlation, for example, to the Spearman and Kendall $\tau$-rank correlation coefficients.

Now we enlist the well-known seven postulates of Renyi (1959), formulated for a measure of dependence $corr\,(X, Y)$ defined in the segment [0, 1]:

(P1) $corr\,(X, Y)$ is defined for any pair of random variables $X$ and $Y$, neither of them being constant with probability 1.

(P2) $corr\,(X, Y) = corr\,(Y, X)$.

(P3) $0 \leq corr\,(X, Y) \leq 1$.

(P4) $corr\,(X, Y) = 0$ if and only if $X$ and $Y$ are independent.

(P5) $corr\,(X, Y) = 1$ if there is a strict dependence between $X$ and $Y$, that is, either $X = g(Y)$ or $Y = h(X)$, where $g(x)$ and $h(X)$ are Borel-measurable functions.

(P6) If the Borel-measurable functions $g(x)$ and $h(x)$ map the real axis in a one-to-one way on to itself, $corr\,(g(X), h(Y)) = corr\,(X, Y)$.

(P7) If the joint distribution of $X$ and $Y$ is normal, then $corr\,(X, Y) = |\rho(X, Y)|$, where $\rho(X, Y)$ is Pearson's product-moment correlation coefficient.

This set of postulates is more restrictive than the proposed set (R1)–(R5) mostly because of the chosen range [0, 1] and the last postulate (P7) that yield the absolute value of Pearson's correlation $|\rho|$. Later we return to this set when considering informational measures of correlation. Moreover, in what follows, we generally focus on the conventional tools of correlation analysis based on Pearson's correlation coefficient and closely related to it measures, implicitly using Renyi's postulates.

## 2.2 Pearson's Correlation Coefficient: Definitions and Interpretations

Below we represent a series of different conceptual and computational definitions of the population and sample Pearson's correlation coefficients $\rho$ and $r$, respectively. Each definition indicates a different way of thinking about this measure within different statistical contexts by using algebraic, geometric, and trigonometric settings (Rogers and Nicewander 1988).

### 2.2.1 Introductory remarks

The traditional for introductory statistics textbooks definitions (1.1) and (1.2) for Pearson's $\rho$ and $r$ can be evidently rewritten as

$$\rho = \frac{cov(X, Y)}{\sigma_X \, \sigma_Y}$$

and

$$r = \frac{1}{n} \sum_{i=1}^{n} \frac{x_i - \overline{x}}{s_x} \, \frac{y_i - \overline{y}}{s_y}, \tag{2.1}$$

where $s_x$ and $s_y$ are the mean squared errors.

Equation (2.1) for the sample correlation coefficient $r$ can be regarded as the sample covariance of the standardized random variables, namely $(X - \mu_X)/\sigma_X$ and $(Y - \mu_Y)/\sigma_Y$.

Pearson's correlation coefficient possesses the properties (R1) and (R2) with the bounds $-1 \le \rho, r \le 1$: the cases $|\rho| = 1$ and $|r| = 1$ correspond to the linear dependence between variables.

Thus, Pearson's correlation coefficient is a measure of the linear interrelationship between random variables. Furthermore, the relations $\rho = 0$ and $r = 0$ do not induce independence of random variables. The typical shapes of correlated data clusters are exhibited in Figs 2.1 to 2.5.

### 2.2.2 Correlation via regression

The problem of estimation of the correlation coefficient $\rho$ is directly related to the linear regression problem of fitting the straight line of the conditional expectation

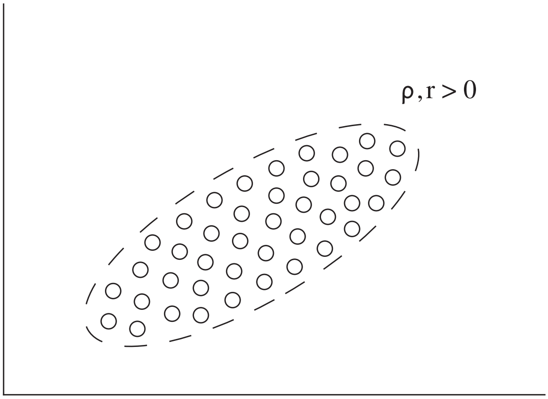


*Figure 2.1 Data with positive correlation.*